APLICACION DE MINERIA DE DATOS PARA LA EXPLORACION Y DETECCION DE PATRONES DELICTIVOS EN ARGENTINA

F. Valenga[1], I. Perversi[2], E. Fernández[2,3], H. Merlino[2,3], D. Rodríguez[2], P. Britos[2,3] y R. García-Martínez[2,3]

[1] Licenciatura en Informática. Universidad de Morón.
[2] Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. ITBA.
[3] Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires.
Argentina
rgm@itba.edu.ar

Resumen

El presente trabajo describe un Proyecto de Minería de Datos en el ámbito de la información criminal, analizando los homicidios dolosos cometidos en la República Argentina mediante una herramienta de distribución libre.

Palabras claves: Minería de Datos. Inteligencia Criminal.

Abstract

This article describes a Project of Data mining in thee area of the criminal information, analyzing. The fraudulent homicides committed in the Republic Argentina using a tool of free distribution.

Keywords: Data Mining. Criminal Intelligence.

1. Estadistica y Mineria de Datos: Abordajes complementarios

El abordaje metodológico-estadístico utilizado por el análisis e interpretación sobre criminalidad en la Argentina actualmente en uso en la Dirección Nacional de Política Criminal (DNPC) es consistente con la tradición científica en el área [Kumar,1996; Marczyk, *et al*.,2005; Creswell, 2003] y con las metodologías utilizadas a nivel mundial en el área [Chen *et al*., 2004; Zeleznikow, 2005; Coplink; 2007] .
La minería de datos, así como el descubrimiento de conocimientos en los datos, integra desarrollos y concepciones provenientes de la estadística, el aprendizaje automático, la visualización de datos y la teoría de bases de datos. Esta fusión de disciplinas muy diversas ha estado motivada (entre otras) por el significativo incremento del volumen de los datos en todas las esferas de la actividad humana y en este caso particular en la necesidad de disponer de la mayor cantidad de elementos para establecer políticas de inteligencia criminal mas ajustadas con base en los datos disponibles en los diferentes soportes.
Ambos abordajes han mostrado ser complementarios. Mientras que la Estadística plantea hipótesis que deben ser validadas a partir de los datos disponibles, la Minería de Datos descubre patrones en los datos disponibles que mediante la interpretación de expertos del dominio propone patrones de comportamiento social (en nuestro caso) no previstos desde el otro abordaje.
En este contexto la Minería de Datos emerge como el siguiente paso evolutivo en el proceso de análisis de datos criminales.
Para validar la utilidad del uso de minería de datos en la exploración y detección de patrones delictivos y su complementariedad con el abordaje estadístico utilizado en la DNPC se han hecho algunos trabajos exploratorios cuyos resultados se presentan en las siguientes secciones.

2. Estado de la Cuestión

A partir de la crisis de finales de 2001, Argentina se vio afectada por una creciente ola de inseguridad caracterizada por un aumento en los índices delictivos y los niveles de violencia. Esta situación fue más profunda en los principales centros urbanos y llevó a tomar acciones coordinadas a nivel nacional tendientes a prevenir el delito. Una de estas medidas fue la creación del Sistema de Alerta Temprana (SAT) por parte del Ministerio de Justicia y Derechos Humanos. En el plano internacional, los ataques terroristas del 11 de septiembre han aumentado significativamente la preocupación por la seguridad interna en todo el mundo. Agencias de inteligencia como la CIA o el FBI procesan y analizan información activamente en busca de actividad terrorista [Chen *et al*., 2004].
En este contexto, el análisis de los registros criminales es fundamental en la prevención del delito. Entre otras cosas, porque permite el diseño de políticas y planes de prevención efectivos. En Argentina este tipo de análisis se ha realizado históricamente mediante herramientas estadísticas descriptivas o deductivas, considerando fundamentalmente variables y relaciones primarias. Sin embargo, muchas veces la estadística descriptiva clásica no refleja la verdadera interrelación de las variables y por lo tanto, el problema real. Este contexto requiere un tratamiento estadístico más complejo que nos obliga a evolucionar en el análisis de información criminal.
En general, el tamaño de las bases de datos está basado en aspectos como la capacidad y eficiencia de almacenamiento y no en su posterior uso o análisis. Por esta razón, en muchos casos, los registros almacenados son demasiado grandes o complejos como para analizar [Kantardzic, 2002] y superan el alcance de la estadística [Hand, 1997]. La Minería de Datos (*Data Mining*) es un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos que busca generar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas [Britos *et al*, 2005].

En el caso de la inteligencia criminal, la gran cantidad de información y de variables intervinientes justifican el uso de herramientas más potentes que la estadística convencional que permitan determinar relaciones multivariantes subyacentes. La minería de datos aplicada a la inteligencia criminal es un campo bastante nuevo y ha tenido un gran impulso en los últimos años, sobre todo en EEUU [Chen *et al*., 2004].

2.1. Tratamiento de la Información Criminal en el mundo

A continuación se describen algunas de las principales experiencias de aplicación de minería de datos en el análisis de información criminal a nivel mundial:

- *Proyecto COPLINK*

  El Proyecto COPLINK fue creado en el año 1997 en el Laboratorio de Inteligencia Artificial de la Universidad de Arizona, en Tucson, con el objetivo de servir de modelo para ser llevado a nivel nacional. Recientemente se ha desarrollado la versión comercial, denominada *COPLINK Solution Suite* [Coplink, 2007].

  Coplink está compuesto por dos sistemas integrados: Coplink Connect y Coplink Detect. El primero busca compartir información criminal entre distintos departamentos policiales, mediante un fácil acceso y una interfase sencilla, integrando distintas fuentes de información. El segundo esta diseñado para detectar de forma automática distintos tipos de asociaciones entre las bases de datos mediante técnicas de minería de datos. Ambos sistemas presentan una interfase visual amigable [Chen et al., 2004]. Entre otras aplicaciones Coplink provee *Análisis de Redes Criminales* [Chen et al., 2004], la cual consiste en: identificar las redes o bandas criminales, sus líderes o integrantes clave y como se relacionan entre sí. En primer lugar se utiliza la técnica de *concept space* para extraer relaciones de los sumarios policiales y construir una posible red de sospechosos. La fuerza del vínculo entre dos sospechosos se mide en base a la frecuencia de hechos en los que participaron ambos. Luego se utiliza *clustering* jerárquico para partir la red en subgrupos y *block modeling* para identificar patrones de interacción entre los mismos. Finalmente se calcula el baricentro de cada subgrupo para determinar su miembro clave o líder.

- *Proyecto OVER*

  El Proyecto OVER comenzó en el año 2000 en Reino Unido como una iniciativa conjunta de la Policía de West Midlands y el Centro de Sistemas de Adaptación y División de Psicología de la Universidad de Sunderland. El proyecto esta enfocado en los casos de robo a domicilio particulares. Sus principales objetivos son [Zelzenikow, 2005]:

  - identificar los recursos críticos para establecer estrategias de prevención y detección más eficientes;
  - proveer de fundamentos empíricos para el desarrollo de planes interdepartamentales orientados a la reducción del delito;
  - identificar la información relevante a ser recolectada en el lugar del hecho, redundando en mejoras de eficiencia y reducción de tiempo del personal policial;
  - alimentar al sistema tanto con información *hard* (información forense) como *soft* (información sobre la escena del delito);
  - analizar la distribución espacio-temporal de los hechos y confirmar las suposiciones sobre tendencias y patrones.

- *Otras proyectos*

- El Departamento de Policía de Ámsterdam utiliza el software de minería de datos DataDetective [Sentient, 2007] junto con Mapinfo para el análisis de registros criminales. Las principales técnicas empleadas son árboles de decisión y redes neuronales de backpropagation. Han unificado varias bases de datos policiales junto con información externa (clima, variables socioeconómicas y demográficas) en un único data warehouse. Los principales usos son:
  - identificación de las causas del comportamiento criminal (por ejemplo casos de reincidencia);
  - identificación de las causas del delito en un determinado barrio;
  - agrupamiento de delitos parecidos en *clusters* y su descripción, permitiendo un abordaje más efectivo;
  - identificación de delitos parecidos utilizando algoritmos *fuzzy search*, relacionando casos no resueltos con casos resueltos;
  - identificación de zonas de aumento del delito (por ejemplo se ha utilizado para la localización de equipos preventivos en operativos de búsqueda de armas);
  - evaluación de la performance policial.
- El Departamento de Policía de Richmond (Virginia) ha desarrollado una aplicación para el análisis de información criminal que combina minería de datos, mediante el software Clementine [SPSS, 2007], junto a un entorno visual aportado por Information Builders [IB, 2007] y una iterfase desarrollada por RTI Internacional [RTI, 2007]. El principal objetivo es optimizar la alocación de recursos, en base a una modalidad preactiva y no reactiva. Por ejemplo durante año nuevo se identificaron las zonas que habían tenido un aumento en los casos de heridos de con arma de fuego el año anterior y para la noche se reforzaron exclusivamente esas zonas. El resultado obtenido fue una reducción del 49% en los casos de este tipo con un menor requerimiento de personal policial aproximadamente 50 agentes menos) [SPSS, 2007].
- La Policía Estatal de Illinois adquirió en 2005 un software de minería de datos del compañía RiverGlass Inc. [RiverGlass, 2007] con el objetivo de analizar la información criminal en tiempo real. El campo de aplicación es muy grande y va desde la seguridad marítima en los puertos a la detección de casos de fraude financiero.
- El Departamento de Policía de San Francisco desarrolló junto a IBM la aplicación CrimeMaps, en base a la tecnología DB2 de IBM [IBM, 2007]. Este software permite a los oficiales mediante un simple explorador web buscar un determinado tipo de crimen, realizar análisis de clustering y fijar niveles umbrales de alerta temprana para un determinado delito en una determinada zona de acuerdo a una frecuencia histórica.
- El Departamento de Policía de Nueva York inició en julio de 2005 el Real Time Crime Center [NYC, 2007]. Este ambicioso proyecto tiene como objetivo conformar un enorme data warehouse y cruzar información de todo tipo mediante herramientas de inteligencia de negocios (como Repotnet 1.1 y Accurint Pro) de forma de detectar patrones de comportamiento y asociaciones antes desapercibidos.

2.2. La información criminal en la Argentina

La presente sección esta basada en las reuniones mantenidas con personal de la Dirección Nacional de Política Criminal del Ministerio de Justicia y Derechos Humanos de la Nación.
Se entiende por información criminal a toda aquella información resultante a partir de un presunto delito o de sus componentes (victima, victimario, propiedades, vehículos, etc.) que sea relevante

para la toma de decisiones a posteriori. Ya sea en la prevención, detección y esclarecimiento del delito como en la prosecución de delincuentes, la mejora de procesos judiciales y la creación de nuevas leyes. Según esta definición la mayor fuente de información criminal es el Sistema Penal, entendido como el conjunto de instituciones y procedimientos presentes en el proceso que transita un hecho delictuoso desde que es registrado por el Estado.
Se puede subdividir al Sistema Penal según las distintas instancias en: Sistema Policial, Sistema Judicial y Sistema Penitenciario. Como muestra la Figura 1, una gran cantidad de hechos ingresa por el Sistema Policial, atraviesa el cuello de botella del Sistema Judicial y egresa a través del Sistema Penitenciario.

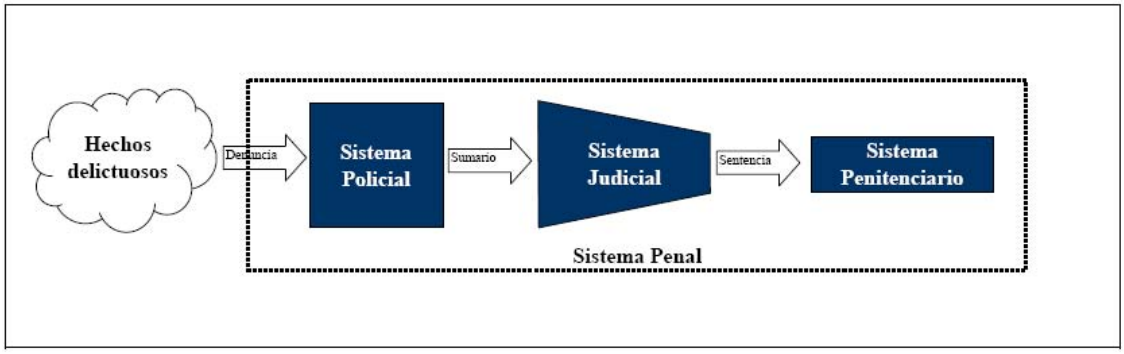

Figura 1- Esquema del Sistema Penal

En Argentina debido al sistema de gobierno federal adoptado, este esquema se replica en cada una de las provincias del territorio nacional.
Este sistema de gobierno tiene fuertes implicancias en la consolidación de la información a nivel nacional. No sólo por la falta de homogeneidad entre las distintas provincias, sino fundamentalmente porque cada provincia tiene autonomía sobre la información generada bajo su jurisdicción y el Estado Nacional no tiene injerencia sobre la misma.
Las funciones de consolidación de información criminal a nivel nacional y confección de la estadística general fueron delegadas, a través de la Ley Nacional 25.266, a la Dirección Nacional de Política Criminal (DNPC) del Ministerio de Justicia y Derechos Humanos de la Nación (MJDHN). Esta ley establece a la DNPC como única fuente oficial de información criminal a nivel nacional. Cabe mencionar que esta capacidad de la DNCP no va en desmedro de que las distintas instituciones del Sistema Penal posean su propia información, e incluso sectores de estadística, para analizar su propia gestión. De hecho, la gran mayoría de las policías nacionales poseen una división de estadística.
Las primeras estadísticas policiales en Argentina comenzaron a ser elaboradas por la Policía de la Capital Federal en 1887 [Blackwelder Y Jonson, 1984]; [Rubial, 1993]; [Sozzo, 2000]. A partir de 1971 toda la información proveniente del Sistema Policial pasó a ser consolidada a nivel nacional por el Registro Nacional de Reincidencia y Estadísticas Criminales (RNREC). Los siguientes 30 años se caracterizaron por estadísticas incompletas, de poca calidad y sin un análisis posterior.
A partir de 1999 se intentó revertir esta situación mediante la creación del Sistema Nacional de Información Criminal (SNIC) y el Sistema de Alerta Temprana (SAT), y la transferencia de las funciones de consolidación y análisis de información criminal a la Dirección Nacional de Política Criminal (DNPC). En julio de 2000 se formalizó esta transferencia mediante la Ley Nacional 25.266 anteriormente mencionada.
El Sistema de Alerta Temprana (SAT) se nutre de cuatro planillas complementarias a la del SNIC que relevan información detallada sobre cuatro aspectos en particular:

- homicidios dolosos;
- homicidios culposos en accidentes de tránsito;
- suicidios;

- delitos contra la propiedad.

En los primeros tres casos se releva información puntual de cada hecho, mientras que el último consta de sumarizaciones parciales según distintas variables. La información relevada en cada caso es la siguiente.

2.3. Introducción a las técnicas de Minería de Datos

La Minería de Datos es un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos [Kantardzic, 2002] que busca generar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas [Britos et al, 2005].
Las principales técnicas de minería de datos se suelen clasificar según su tarea de descubrimiento, en tal sentido a continuación se describen las clasificaciones consideradas más relevantes para el actual proyecto:

- Agrupación o clustering.
- Clasificación.

A continuación se realiza una breve descripción de cada una de estas clasificaciones:

- *Agrupación de Datos o Clustering:*

  La agrupación o el clustering consiste en agrupar un conjunto de datos, sin tener clases predefinidas, basándose en la similitud de los valores de los atributos de los distintos datos. Esta agrupación, a diferencia de la clasificación, se realiza de forma no supervisada, ya que no se conoce de antemano las clases del conjunto de datos de entrenamiento. El clustering identifica clusters, o regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [Chen y Han, 1996]. El clustering se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters [Han y Kamber, 2001].

  K-Means [Britos et al., 2005] en un método particional de clustering donde se construye una partición de una base de datos D de n objetos en un conjunto de k grupos, buscando optimizar el criterio de particionamiento elegido. En K-Means cada grupo está representado por su centro. K-Means intenta formar k grupos, con k predeterminado antes del inicio del proceso. Asume que los atributos de los objetos forman un vector espacial. El objetivo que se intenta alcanzar es minimizar la varianza total intra-grupo o la función de error cuadrático.

- *Clasificación de Datos:*

  La clasificación se utiliza para clasificar un conjunto de datos basado en los valores de sus atributos. Por ejemplo, se podría clasificar a distintas personas para el otorgamiento de un préstamo en riesgo bajo, medio y alto, teniendo en cuenta información histórica de las mismas.

  La clasificación encuentra las propiedades comunes entre un conjunto de objetos y los clasifica en diferentes clases, de acuerdo a un modelo de clasificación. Para construir este modelo, se utiliza un conjunto de entrenamiento, en el que cada instancia consiste en un conjunto de atributos y el valor de la clase a la cual pertenece [Chen y Han, 1996].

  Los algoritmos mayormente utilizados para las tareas de clasificación son los algoritmos de inducción. En la actualidad existen numerosos enfoques de algoritmos de inducción y variedad en cada enfoque, el presente trabajo hará hincapié en aquellos orientados a generar árboles de decisión.

  ID3 es un sistema típico de construcción de árboles de decisión, el cual adopta una estrategia de arriba hacia abajo e inspecciona solo una parte del espacio de búsqueda. ID-3 garantiza que será encontrado un árbol simple, pero no necesariamente el más simple. ID-3 utiliza la teoría de la información para minimizar la cantidad de pruebas para clasificar un objeto. Una heurística selecciona el atributo que provee la mayor ganancia de la información. Una extensión a ID3, C4.5 [Weka, 2007] extiende el dominio de clasificación de atributos categóricos a numéricos. J48 es una implementación mejorada del algoritmo de árboles de decisión C4.5. El algoritmo J48 funciona bien con atributos nominales y numéricos. Un paso importante en la construcción del árbol de decisión es la poda, la cual elimina las ramas no necesarias, resultando en una clasificación más rápida y una mejora en la precisión de la clasificación de datos [Han y Kamber, 2001].

Existen en la actualidad varias herramientas de libre distribución que permiten aplicar las técnicas antes mencionadas, entre ellas se encuentra *Weka* [Weka, 2007]. La cuan fue desarrollada originalmente en la universidad de Waikato y hoy día es accesible fácilmente desde Internet.

2.4. Antecedentes vinculados al tratamiento de la información criminal en la Argentina

En Argentina no se conoce ninguna experiencia de aplicación de minería de datos a información criminal. Sin embargo hay dos proyectos relacionados que merecen ser mencionados.

- *El Proyecto SURC*

  A comienzos de 2004 el entonces Ministerio de Justicia, Seguridad y Derechos Humanos lanzó el proyecto del Sistema Unificado de Registros Criminales (SURC). El objetivo era interconectar y articular las instituciones del Sistema Policial y el Sistema Judicial mediante una red en la cual todos tuvieran acceso a un banco de datos común, alimentado en tiempo real y del cual se pudieran realizar consultas *online*. Este banco de datos contemplaba información diversa [SSI-MI, 2004]:

  - Registro de hechos: características generales del hecho denunciado (lugar, día, hora, delito denunciado y comisaría interviniente).
  - Registro de denunciantes: identidad y características de la víctima o denunciante.
  - Registro de autores identificados: identidad, características, historial criminal e imágenes de los autores.

- Registro de autores no identificados: descripción de los NN (contextura física, edad aproximada, estatura, color de pelo, señas particulares, frases frecuentes, etc.).
- Registro de elementos robados: información útil para la identificación de los objetos robados.
- Registro de autos robados: marca, modelo, color, número de patente, número de motor, características particulares, etc.
- Registro de armas secuestradas: características de las armas secuestradas, vinculando esta base con otros sistemas como el Ibis.
- Registro de evidencias: descripción de huellas y pistas relevadas en la escena del crimen.
- Mapa del delito: presentación de los hechos en forma gráfica y geo-referenciada mediante GIS.

Este proyecto de gran alcance contemplaba una implementación progresiva, comenzando por la Ciudad Autónoma de Buenos Aires y avanzando hacia las provincias. Sin embargo tras la salida del entonces Ministro de Justicia, Seguridad y Derechos Humanos, Gustavo Béliz, y gran parte de su equipo de trabajo, en julio de 2004, el proyecto quedó congelado. Tiempo más tarde se trasladaron las funciones de seguridad de la esfera del Ministerio de Justicia al Ministerio del Interior, y consigo el proyecto SURC. Actualmente el proyecto permanece vigente, con radicación en la Secretaría de Inteligencia Criminal del Ministerio del Interior de la Nación, pero relegado y con un cambio de enfoque respecto al original.

- *El Mapa del Delito de la Ciudad Autónoma de Buenos Aires*

  El Ministerio Público Fiscal de la Nación (MPFN) es una de las pocas instituciones judiciales de Argentina que posee un sistema de información digitalizada. Cuenta con una base de datos de los hechos delictivos de autoría desconocida (NN) registrados en Capital Federal. Esta base contiene información referida al hecho, como ser: tipo de delito, fecha, lugar y cantidad de víctimas.
  Asimismo el Centro de Información Metropolitana (CIM), radicado en la Facultad de Arquitectura, Diseño y Urbanismo (FADU) de la Universidad de Buenos Aires (UBA), posee el Sistema de Información Territorial del Área Metropolitana de Buenos Aires (SAT/AMBA). Este sistema consiste en la base cartográfica digital de todo el AMBA para ser utilizada bajo GISs (*Geographical Information Systems*). No sólo posee los elementos tradicionales (calles, avenidas, vías del ferrocarril, plazas, etc.) sino también la visualización de las demarcaciones zonales (barrios, centros de gestión y participación, comisarías, etc.) y gran parte del equipamiento urbano (escuelas, clubes, bancos, etc.).
  En 2002 ambas instituciones firmaron un "Convenio de Asistencia, complementación y Cooperación" con el objetivo de que el CIM elaborase el Mapa del Delito de la Ciudad Autónoma de Buenos Aires con la información suministrada por el MPFN.
  Si bien la existencia de este mapa no es muy conocida y pese a que cuenta con ciertas limitaciones (únicamente hechos ocurridos en Capital Federal con autoría desconocida), su aporte en el análisis de la situación criminal es muy valioso[Behar, Lucilli, 2003].

3. Definición del problema

El Sistema de Alerta Temprana (SAT) creado por la Dirección Nacional de Política Criminal (DNPC) del Ministerio de Justicia y Derechos Humanos (MJDH) consiste en un relevamiento a nivel nacional de determinados tipos de delitos. En la actualidad esta información es analizada mediante un análisis estadístico básico, sin hacer un aprovechamiento exhaustivo de la información mediante el uso de técnicas o herramientas de Minería de Datos.

4. Solución propuesta

Para aumentar el poder de la información criminal existente, se propone llevar a delante un proyecto de Minería de Datos mediante la herramienta Weka 3.5.3. tomando como base los datos reportados por el sistema SAT.

A continuación se describe la estrategia de aplicación de algoritmos:

1. En primer lugar se prevee aplicar el algoritmo Simple K-means para clusterizar. Estos resultados serán analizados y convalidados con los usuarios, utilizando para ello los informes emitidos por Weka.
2. Una vez estabilizados los cluster se utilizarán algoritmos de Inducción para explicar el comportamiento de los mismos de una forma mas descriptiva. Para ello se utilizará el algoritmo J48.

5. Demostración de la Solución

5.1. Descripción del Dataset

Se analizaron 1810 registros de la base de datos “Homicidios Dolosos” correspondientes a la totalidad de hechos registrados durante 2005, provenientes del SAT. Cuyos atributos se describen a continuación en la tabla 1:

Provincia	Departamento	Día del mes	Mes	Día de la semana
Hora	Lugar	Arma	Otro delito	

Tabla 1- Atributos del Dataset

5.2. Resultados del proceso de Agrupamiento

5.2.1. Centroides

A continuación, en la tabla 2, se describen los centroides obtenidos:

		Atributos categóricos (modas)				Atributos continuos (medias)			
	Cant. (%)	Provincia	Lugar	Arma	Otro Delito	Hora	Día Semana	Día Mes	Mes
Cluster 0	22%	BsAs	Vía Pública	de Fuego	Robo	19	Sábado	16	7
Cluster 1	43%	BsAs	Vía Pública	de Fuego	No Hubo	17	Sábado	15	7
Cluster 2	35%	BsAs	Domicilio Particular	Blanca	No Hubo	21	Sábado	15	7
General	100%	BsAs	Vía Pública	de Fuego	No Hubo	19	Sábado	15	7

Tabla 2- Centroides

5.2.2. Gráficos de Barras

La distribución de los clusters entre las variables de los distintos atributos permite comprender el nivel de significancia de los mismos (ver figura 2). En este caso, si los clusters fueran irrelevantes, esperaríamos encontrar una proporción aproximada de 43% rojo (cluster 1); 22% azul (cluster 0) y 35% turquesa (cluster 2) en cada variable de cada atributo. Si bien en algunos atributos esta

proporción se cumple (*día mes y provincia*) en otros existen interacciones significativas (por ejemplo cluster 2 con *arma blanca* y *domicilio particular*):

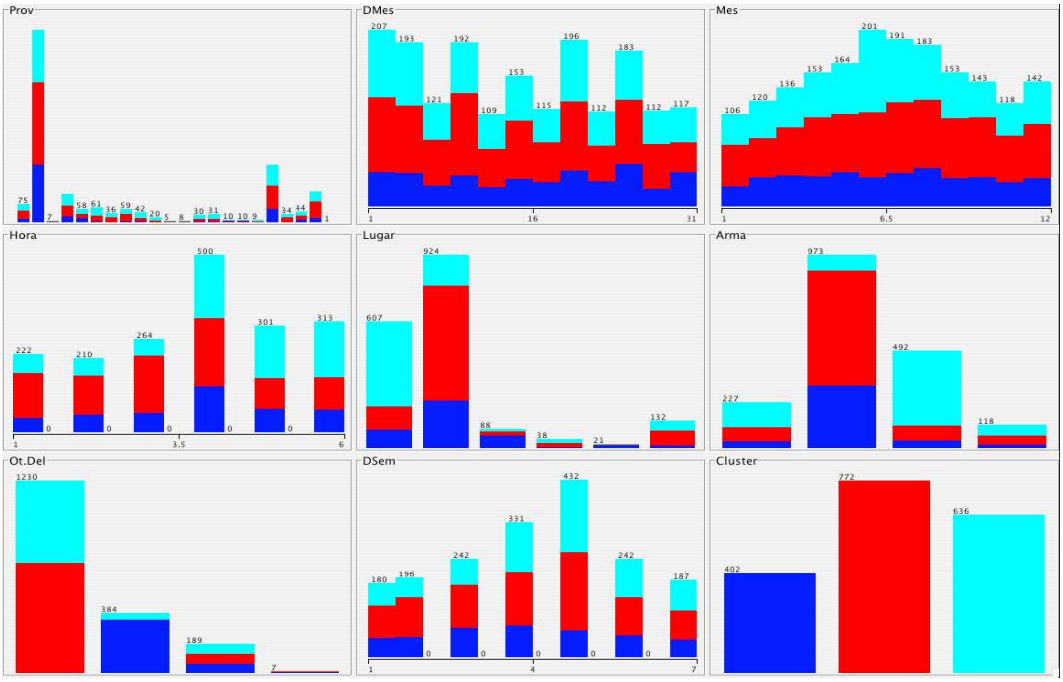

Figura 2- Distribución de Clusters

5.2.3. Gráficos de dispersión

A continuación se describen los cluster en base a dos de los atributos más representativos:

Distribución de los clusters según el atributo lugar (ver figura3) : Mientras el cluster 2 esta muy concentrado en domicilio particular y el cluster 1 en vía pública, el cluster 0 se encuentra distribuido más homogéneamente [Figura 3]. Si bien este último presenta la mayoría de registros en domicilio particular, tiene una alta proporción de homicidios en comercios respecto a los otros clusters.

Distribución de los clusters según el atributo arma (ver figura 4) : El cluster 1 y el cluster 0 presentan una distribución similar, con una alta concentración en arma de fuego, seguida por arma blanca y prácticamente muy pocos casos sin arma [Figura 4]. En contraposición el cluster 2 presenta muy pocos casos con arma de fuego (una proporción muy baja respecto a la proporción global) y muchos casos sin arma (una proporción muy alta respecto a la proporción global).

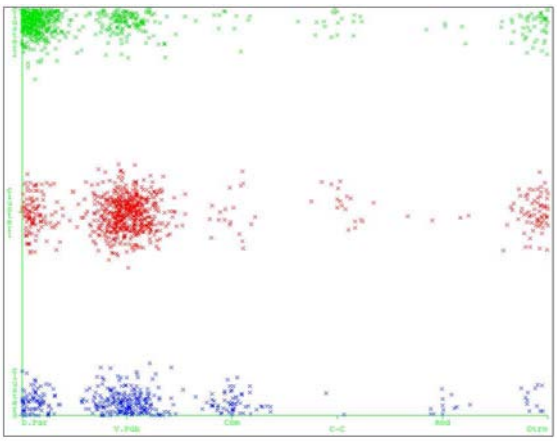

Figura 3- Distribución según atributo lugar

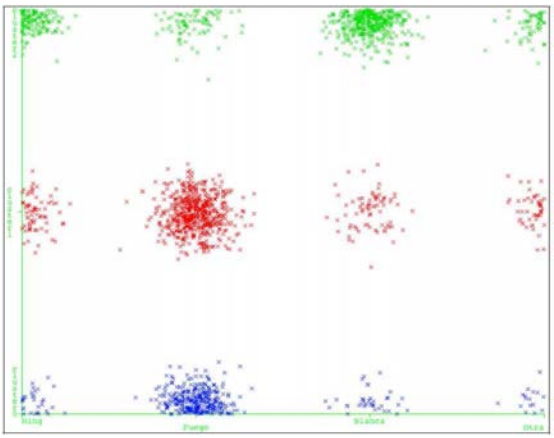

Figura 4- Distribución según atributo Arma

5.2.4. Análisis preliminar de los cluster

En base a la información que surge de este análisis podemos dar una primera interpretación a los clusters:

- *Cluster 0 (22%): esta caracterizado por homicidios mayoritariamente en ocasión de robo y con arma de fuego. En principio diremos que se trata de "homicidios en ocasión de robo".*
- *Cluster 1 (43%): es el que más registros agrupa y el más parecido a la media global. Está caracterizado por homicidios mayoritariamente en la vía pública con arma de fuego y sin la existencia de otro delito. Se podrían interpretar como "homicidios en ocasión de riña o ajuste de cuentas".*
- *Cluster 2 (35%): es el más particular de los clusters, ya que la mayoría de sus registros presentan casos de homicidios sin arma de fuego y en domicilio particular. Los denominaremos "homicidios en ocasión de emoción violenta".*

5.3. Árbol de clasificación

Como el árbol clasificatorio obtenido es muy grande, a continuación, en la tabla 3, se describe el 9% de las reglas de clasificación extraídas del árbol, las cuales clasifican el 66% de las instancias (1200). Estas reglas son las siguientes:

Regla 1 SI otro delito = no hubo Y arma = fuego Y lugar = V.Púb. ENTONCES Cluster 1 (362)	**Regla 2** SI otro delito = robo Y arma = fuego ENTONCES Cluster 0 (272)	**Regla 3** SI otro delito = no hubo Y arma = blanca Y lugar = D.Part. ENTONCES Cluster 2 (134)
Regla 4 SI otro delito = no hubo Y arma = blanca Y lugar = V.Pub. Y DSem = Sa-Ma Y hora = 19-8 ENTONCES Cluster 2 (109)	**Regla 5** SI otro delito = no hubo Y arma = ninguna Y lugar = D.Part. ENTONCES Cluster 2 (87)	**Regla 6** SI otro delito = no hubo Y arma = fuego Y lugar = D.Part. Y hora = 8-19 ENTONCES Cluster 1 (85/3)
Regla 7 SI otro delito = no hubo Y arma = blanca Y lugar = V.Pub. Y hora = 8-16 ENTONCES Cluster 1 (55)	**Regla 8** SI otro delito = no hubo Y arma = ninguna Y lugar = V.Pub. ENTONCES Cluster 1 (48)	**Regla 9** SI otro delito = no hubo Y arma = fuego Y lugar D.Part. Y hora = 20-8 ENTONCES Cluster 1 (48)

Tabla 3- Reglas de Clasificación

Estas reglas fueron consultadas con los especialistas y permitieron confirmar la interpretación hecha anteriormente. Al respecto, los especialistas comentaron que hasta el momento ellos solían clasificar a los homicidios en dos grupos, según el vínculo existente entre la víctima y el agresor:

- *los casos de robo, en los que víctima y agresor no se conocen;*
- *el resto de los casos, denominados "homicidios en conflictos interpersonales".*

La conclusión arribada junto con los especialistas es que se trata de **dos tipos de conflictos interpersonales distintos**, uno más bien **familiar** (dentro del domicilio particular) e **impulsivo** (sin

arma) y otro más bien **vecinal** o de **ajuste de cuentas** (vía pública) y con cierto nivel de **premeditación o pre-intencionalidad** (arma de fuego).
En el medio de estos dos grupos extremos están los casos de armas blancas, difíciles de asignar a priori a una u otra modalidad

6. Conclusiones

El presente trabajo ha demostrado no sólo que es factible aplicar minería de datos a la información criminal en Argentina, sino también su alto valor agregado para el análisis y la generación de nuevo conocimiento.
La experiencia realizada en conjunto con la Dirección Nacional de Política Criminal (DNPC) del Ministerio de Justicia y Derechos Humanos de la Nación (MJDHN), basa la factibilidad en los siguientes puntos:

- *existe gran cantidad de información que actualmente no esta siendo aprovechada en toda su dimensión;*
- *existe un software de minería de datos de distribución libre y gratuita, fácil de usar y que contiene las herramientas necesarias para el análisis;*
- *este software de minería de datos puede ser utilizado por una persona ajena al ámbito informático con una capacitación básica.*

Los resultados experimentales obtenidos han sido validados por los especialistas de la DNPC. Estos resultados han permitido tanto confirmar conceptos preexistentes (pero con una justificación sustentada en los datos), como generar nuevas piezas de conocimiento. Al respecto se han identificado tres patrones distintos de homicidios dolosos en base a los hechos ocurridos en Argentina durante 2005.

7. Futuras líneas de investigación

En primer lugar se propone aumentar el alcance de la información de la DNPC a ser analizada con este tipo de técnicas. Esto implica tanto una expansión transversal, haciendo uso de otras bases de datos como la de “homicidios culposos en accidentes de tránsito”; como longitudinal, analizando la información histórica existente para detectar patrones de evolución temporal en cuanto a las modalidades delictivas.
En segundo lugar se sugiere el diseño de procedimientos estándar de minería de datos con *Weka* para ser implementados en la DNPC. Esta batería de procedimientos les permitiría a los analistas de la DNPC extraer e identificar patrones y asociaciones en forma automatizada y estandarizada.
En tercer lugar se propone proceder al análisis de la información geográfica relevada por la DNPC (que hoy no es aprovechada) mediante GISs (*Geographical Information Systems*). Este tipo de análisis permitiría detectar, por ejemplo, zonas de alta densidad de homicidios en accidentes de tránsito.
Finalmente se propone expandir el uso de estas técnicas a las fuerzas de seguridad, en donde estas aplicaciones han encontrado su mayor aplicación a nivel mundial.

8. Referencias

Behar, A. M., P. Lucilli, 2003. *Mapa del delito de la Ciudad Autónoma de Buenos Aires*. Terceras Jornadas de Jóvenes Investigadores, Instituto Gino Germani.

Blackwelder, J.K., L.L. Jonson, 1984. *Estadística Criminal y Acción Policial en Buenos Aires, 1887-1914*. Desarrollo Económico, 93, Vol. 24, 1984, pp. 109-122.

Britos, P., Hossian, A., García-Martinez, R. y Sierra, E. 2005. Minería de Datos Basada en Sistemas Inteligentes. Editorial Nueva Librería.. ISBN 987-1104-30-8 .

Chen, H., W. Chung, J. Xu, G. Wang, Y. Qin, M. Chau, 2004. *Crime Data Mining: A General Framework and Some Examples*. IEEE Computer Society, vol. 37, no. 4, pp. 50-56.

Chen, M., Han, J., *Data mining: An overview from database perspective*. IEEE Transactions on Knowledge and Data Eng., 1996.

Coplink, 2007. COPLINK Solution Suite. www.coplink.com. Acceso mayo 2007.

Creswell, J. 2003. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage Publications.

Han, J., Kamber, M.; *Data mining: Concepts and techniques*. Morgan Kauffmann Publishers, 2001.

Hand, D. J., 1997. *Data Mining: Statistics and More?.* The American Statistician.

IB, 2007. Information Builders. http://www.informationbuilders.com. Acceso mayo 2007.

IBM, 2007. Internacional Business Machines. http://www03.ibm.com/industries/goverment/doc/content/news/pressrelease/1019264109.html Acceso mayo 2007.

Kantardzic, M. 2002. *Data Mining: Concepts, models, methods and algorithms.* Wiley-IEEE Press. ISBN 0-471-22852-4.

Kumar, R.; 1996. *Research Methodology: A Step-by-Step Guide for Beginners*. Addison Wesley.

Marczyk, G., DeMatteo, D., Festinger, D.; 2005. *Essentials of Research Design and Methodology* (Essentials of Behavioral Science). John Wiley & Sons.

NYC, 2007. *New York Police Department Real Time Crime Center*. http://www.nyc.gov/html/nypd/html/dcpi/RTCCRevisedFINALWEB_files/frame.htmAcceso mayo 2007.

Quinlan, J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

RiverGlass, 2007. RiverGlass Inc. http://www.riverglassinc.com/solutions/intelligence.php Acceso mayo 2007.

RTI, 2007. Research Triangle Institute. http://www.rti.org . Acceso mayo 2007.

Rubial B.C., 1993. *Ideología del Control Social, 1880-1920*. Centro Editor de América Latina, Buenos Aires, Argentina.

Sentient, 2007. Sentient Information Systems. http://www.sentient.nl. Acceso mayo 2007.

Sozzo, M., 2000. *Pintando a Través de Números: Fuentes Estadísticas de Conocimiento y Gobierno Democrático de la Cuestión Criminal en Argentina*. http://www.ilsed.org/index.php?option=com_docman&task=doc_view&gid=159&itemid=44 Acceso mayo 2007.

SPSS, 2007. SPSS Inc. URL:http://www.spss.com/success/pdf/CS%20%20Richmond % 20 PD%20LR.pdf. Acceso mayo 2007.

SSI-MI, 2004. *Presentación Institucional Proyecto SURC*. Secretaría de Seguridad Interior, Ministerio del Interior de la República Argentina.

Weka, 2007. *Data Mining Software in Java*; http://www.cs.waikato.ac.nz/ml/weka/ Acceso mayo 2007.

Zeleznikow, J., 2005. *Using Data Mining to Detect Criminal Networks*. www.aic.gov.au/conferences/occasional/2005-04.zeleznikow.html. Acceso mayo 2007.